Classifying the Wine Quality with the Taste of Machine Learning Models

Northwood University

Prof: Mighty Itauma Itauma

Created By: Sai Srujana Jakkala

Smita Karande

ABSTRACT *-- Wine, sounds simple yet has complex significance in the world. It is not just a glass of wine but creates friendships, memories, relationships, partnerships & much more. As said "Wine makes every meal an occasion, every table more elegant, every day more civilized." — André Simon.*

Why not we think more civilized to serve the best wine to the world with the taste of Machine learning. In this paper we are focusing on wine quality as an important factor for both producers and consumers. The quality of wine depends on many properties and chemical composition. Our aim is to classify the quality of wine based on its various chemical compositions. The dataset has various features like acidity, citric acid, residual sugar, chlorides, density, PH value and many more.

We are focusing on developing a classification model to categorize wine quality as good or bad based on the 0 to 10 scaling. This modeling includes machine learning algorithms like logistic regression, Random Forest, support vector model and other modeling techniques. Model performance will be evaluated using F1 score, precision, recall and accuracy values to find the best model for classifying the wine quality. This study aims to support the winemakers in assessing the quality in prior and serve the best quality wine to the consumers.

Keywords: **Machine learning, Classification, Chemical compositions, modeling techniques, Model performance, F1 Score, precision, accuracy**

I. INTRODUCTION

Wine is the most consumed beverage by all the age groups across the globe. The quality of wine is important for both producers and consumers. The quality assessment and certified wines are important to increase the business revenue by the wine producers. Traditionally, the wines are assessed using the wine tasting methods at the end of production cycle but with changing preferences of each person, it becomes difficult to assess the quality of wines based on tasting methods. As the technology advanced the producers use multiple devices during the different production phases to assess the quality of wines that could save lot of time and money as they can realize the quality at early production stages rather than restarting the production in case the wine disqualifies the quality assessment.

As the devices assess the quality, they also capture the data with different chemical compositions involved in wine production and the quality of wines being produced using those composition levels. With the boom of Artificial intelligence (AI) and Machine learning (ML) being the subset of AI helps in performing various models in determining the quality from the available data of its chemical compositions. Here the features are tuned or controlled which directly contributes to quality of wines. By this the producers get better idea of composition values that can be used at early development phases that result in good quality of wines. Also help in building different brands of wines based on the quality strength.

This Machine learning paper is about wine quality classification model using a combined dataset of both Red and white wines from UCI machine learning repository where the target variable is 'Quality'. The data used in this dataset is from Portugal Vinho Verde wines. Our aim is to classify Wine quality influenced by its chemical compositions like acidity, alcohol, density, chlorides, PH etc., using machine learning techniques. The quality of the wines is scaled from 0 to 10 with 0 being very bad to 10 being very good based on its chemical compositions used in different wines. This quality assessment helps both producers prepare, and consumers take good quality of wine. The producers can supply good quality certified wines to the consumer market.

The datasets can be used to perform both classification and regression methods based on the treatment of target variable Quality. Classification method is a favorable technique if target variable is discreet whereas regression is favorable if target variable is continuous.

II. LITERATURE REVIEW

Wine quality heavily rely on machine learning techniques during the production. As we researched on different wine quality datasets and machine learning techniques, studied few research papers that gave me an overview of modelling techniques. Most of the research was based on red wine quality and its prediction using various chemical compositions. This would walk through the literature surveys of few research papers that gave us an idea of data processing, model selection and model evaluation for the wine datasets.

Bhardwaj et al.[14] analyzed the chemical compositions from the New Zeeland based wines. Different types of feature

selection techniques were analyzed, and the findings were compared. Attributes found to be useful from different models were selected to predict the wine quality. Also, multiple machine learning techniques were trained and tested. This paper uses two different sets of 18 Pinto Noir wines evaluated by the experts on the quality. To predict the wine quality, they developed a framework that uses machine learning techniques using huge datasets.

Gupta et al.'s11 research paper uses the key factor required for predicting the wine quality using machine learning techniques. The use of different models like Radom Forest, KNN etc. These models help in assessment of wine quality as average, bad or good quality of wines. The goal was to determine the quality of red wines using its wide range of characteristics. Some of the machine learning techniques like Random Forest, Naïve bayes and SVM methods are performed, and the results are compared on the training and test sets. Many performances are computed, and the best three models is predicted based on the results as to which algorithm predicts the high accuracy results. In this paper the random forest model outperforms with an accuracy of 81.96% and was considered as the best model.

III. METHODOLOGY

This dataset and analysis were performed in Jupiter notebook on VS code platform, Python 3.11.7 version using 'sk learn' packages and other libraries to execute the data.

A. About Dataset:

The dataset is a merged data of two different variants red and white wines with new column added that classifies the wine variant. This data is from UCI machine learning repository that assess the quality of the wines using various chemical compositions as features or attributes and "Quality" as a target or dependent variable. The quality ranges from 0 as very bad to 10 as very good.
The dataset contains total of 6,497 wines including both red and white wines. The independent variables include type (red, white), fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free Sulphur dioxide, total Sulphur dioxide, density, PH, sulphates, alcohol and dependent variable quality (0 to 10).

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Table 1: Shows the Variables from the Dataset

B. Data overview:

The data can be solved using regression or classification technique. The overall data is imbalanced with many normal wines than poor and excellent wines. These can be tested using outliers. Some of the input variables are irrelevant and this could be tested using feature selection method.

There are two datatypes where 'type' is an object and other features are float. There are no missing values in the dataset. The overall summary talks about the distribution of various features like average alcohol content is about 10.

```
Data Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6497 non-null   float64
 2   volatile acidity      6497 non-null   float64
 3   citric acid           6497 non-null   float64
 4   residual sugar        6497 non-null   float64
 5   chlorides             6497 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6497 non-null   float64
 10  sulphates             6497 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
None
```

Fig:1 Summary Statistics

To favor the classification task, we classified the wine quality into different categories of low, medium and high. The low-quality wines range between 0 to 5, Medium quality range between 6 and 7 whereas high quality of wine is between 8 to 10. These categories will help us perform effective classification task.

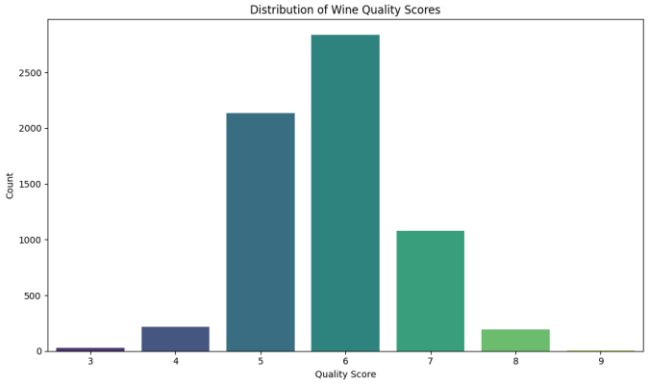


Fig 2: Distribution of Wine Quality Scores

C. Exploratory Data Analysis:

Here, we are trying to analyze the correlation between the different features with the dependent variable quality to identify the most important features. By generating a correlation heatmap, alcohol content is the most important feature composition in determining the wine quality. Other important factors include citric acid, free Sulphur dioxide, sulphates and PH level.

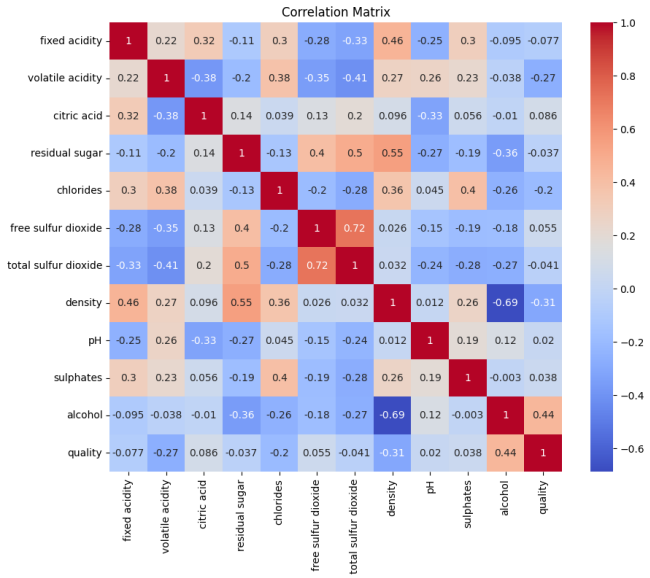


Fig:3: Correlation Matrix of Features

The correlation matrix shows the relationship between different features and the wine quality:

* Alcohol has the strongest positive correlation (0.44) with quality, indicating that higher alcohol content is associated with higher-quality wines.
* Density has the strongest negative correlation (-0.31) with quality, suggesting that lower-density wines tend to be of higher quality.
* Volatile acidity has a moderate negative correlation (-0.27) with quality, implying that lower volatile acidity is associated with higher-quality wines.
* Chlorides also show a negative correlation (-0.20) with quality.

The heatmap visualizes the correlations between all features. The darker red colors indicate strong positive correlations, while darker blue colors indicate strong negative correlations.

Based on the correlation analysis, the most important features impacting wine quality appear to be
a. Alcohol (positive impact)
b. Density (negative impact)
c. Volatile acidity (negative impact)
d. Chlorides (negative impact)
e. Citric acid (slight positive impact)
f. Free sulfur dioxide (slight positive impact)

```
Top 5 features correlated with quality:
alcohol                0.444319
citric acid            0.085532
free sulfur dioxide    0.055463
sulphates              0.038485
pH                     0.019506
Name: quality, dtype: float64
```

Fig:4: Top 5 Features Correlated with Quality:

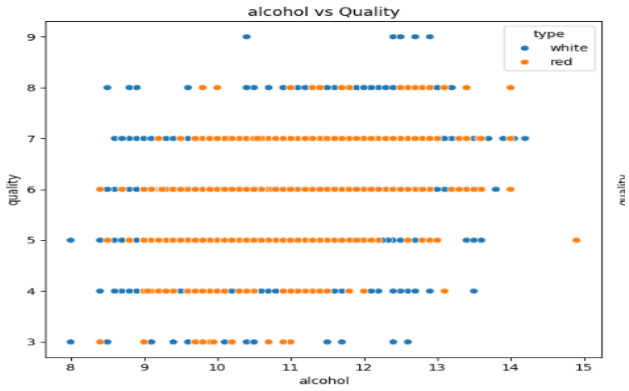


Fig 5: Alcohol Vs Quality

D. Data Preprocessing:

1. Handling missing values: there were no missing values in the data indicating no implication required.
2. Conversion of categorical variable: We converted the ‘type’ variable to binary values.
3. Feature Scaling: The machine learning model performs better with normalized data. To achieve better accuracy, feature scaling was applied to input variables. The ‘StandardScaler’ from sklearn. preprocessing was used for data scaling, so they followed a normal distribution. The scaling was done for training and test data to follow consistency in model building.
4. Splitting data into Train-Test sets: For model performance evaluation data was split into a training set with data of 80% and a test set with data of 20%. This ensures that the model can be tested on unseen data.

E. Model Training and Evaluation;

In this step, multiple classification models were used to predict the quality. This includes simple and complex modeling techniques and performance was compared. We used Logistic regression, Decision tree classifier, Random Forest classifier, Gradient boosting classifier, Support vector machine(SVM), K-nearest neighbor, and Naïve Bayes classifier models.

```
Model Comparison:
                 Model  Accuracy  Precision    Recall
0  Logistic Regression  0.722308   0.701273  0.722308
1        Decision Tree  0.750769   0.762363  0.750769
2    Gradient Boosting  0.762308   0.748656  0.762308
3                  SVM  0.764615   0.743567  0.764615
4                  KNN  0.723846   0.716138  0.723846
5        Random Forest  0.825385   0.828275  0.825385
6          Naive Bayes  0.663077   0.662254  0.663077
```

Fig: 6 Model Comparision

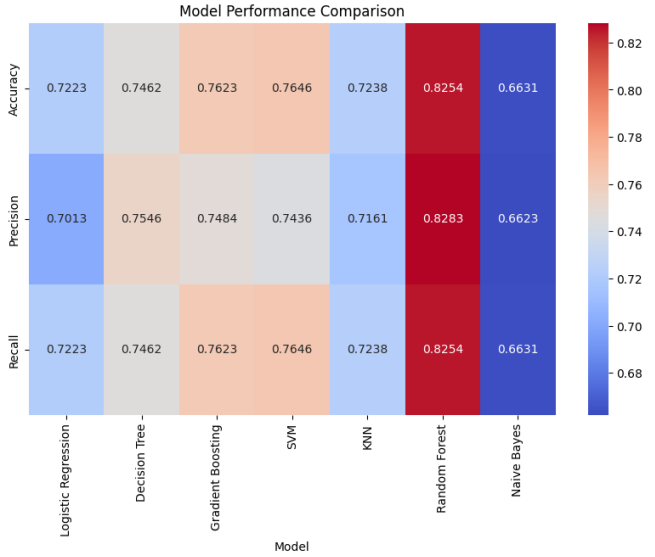


Fig: 7 Model Comaprision Matrix

- **Accuracy**: The proportion of correctly classified instances out of the total instances
- **Precision**: The proportion of correctly predicted positive instances out of all instances predicted as positive.
- **Recall (Sensitivity)**: The proportion of correctly predicted positive instances out of all actual positive instances.

Each model was trained using (X_train_scaled, y_train) using scaled training data. The training was done on regular parameters, but the tunings were performed when necessary. The classification report is generated that compares all the model's performance using the evaluation metrics.

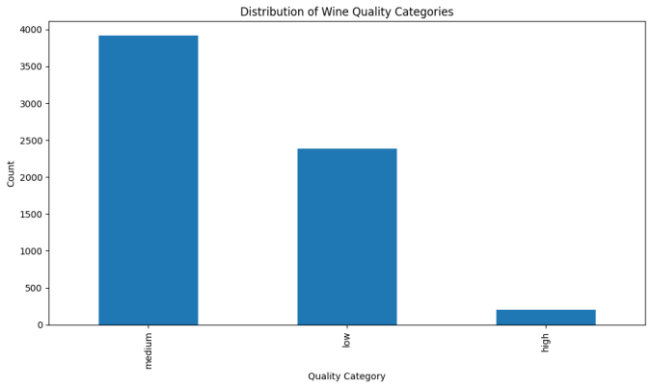


Fig: 8 Distribution of Wine Quality Categories

The wine quality scores are categorized into three distinct groups: low, medium, and high. These categories help in simplifying the classification task and making the analysis more interpretable.

- **Low Quality Wines:** Wines with a quality score of 5 or below.
- **Medium Quality Wines:** Quality Score between 6 and 7 ,
- **High Quality Score:** Quality score more than 8.

As we can see in the given Bar chart the medium quality category has the larger number of wines.

1. Tuning:

"*Some models required tuning for example in logistic regression (max _iter) parameter was increased to 1000 for convergence and Random Forest, the total number of trees (n_estimators) was equalized to 100 for performance balance.*"

F. Model Performance & Results:

By comparing all the models, we can understand that Random Forest model is the best performing model with accuracy of about 83% indicating that its classifying wine quality into different categories(low, medium, high) quiet well. By checking all the metrics in this model

1. Best Performing model

```
Model Performance Metrics:
Accuracy: 0.8254
Precision: 0.8283
Recall: 0.8254
Detailed Classification Report:
              precision    recall  f1-score   support

        high       1.00      0.25      0.40        32
         low       0.78      0.78      0.78       468
      medium       0.85      0.87      0.86       800

    accuracy                           0.83      1300
   macro avg       0.88      0.64      0.68      1300
weighted avg       0.83      0.83      0.82      1300
```

Fig: 9 : Model Performance Metrics

Accuracy: 0.8254 (82.54%) - This means that our model correctly predicts the quality category for about 82.54% of the wines in the test set.
Precision: 0.8283 (82.83%) - This is the weighted average precision across all categories. It indicates that when our model predicts a certain quality category, it's correct about 82.83% of the time.
Recall: 0.8254 (82.54%) - This is the weighted average recall across all categories. It means that our model correctly identifies 82.54% of the wines belonging to each category.

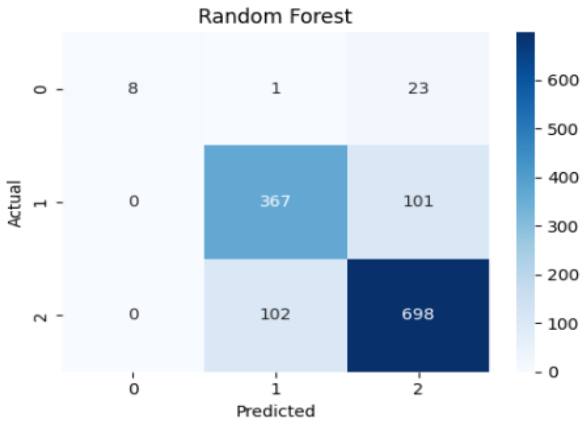


Fig:10: Confusion Matrix for Random Forest

2. **Confusion Matrix:**

A table that provides a more detailed breakdown of the classification results, showing the counts of true positives, true negatives, false positives, and false negatives.

By analyzing the different categories:

- **High-quality wines:** Perfect precision (1.00) but low recall (0.25) which means the model rarely misclassifies other wines as high quality, but it misses many truly high-quality wines.
- **Low-quality wines:** Balanced precision and recall (both 0.78) meaning that the model is consistent in identifying low-quality wines.
- **Medium quality wines:** High precision (0.85) and recall (0.87) indicating that the model performs best on this category, which is also the most common.

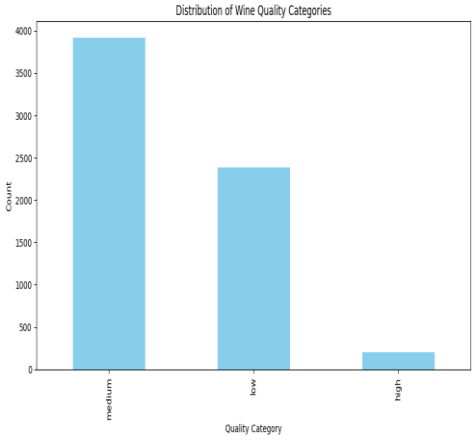
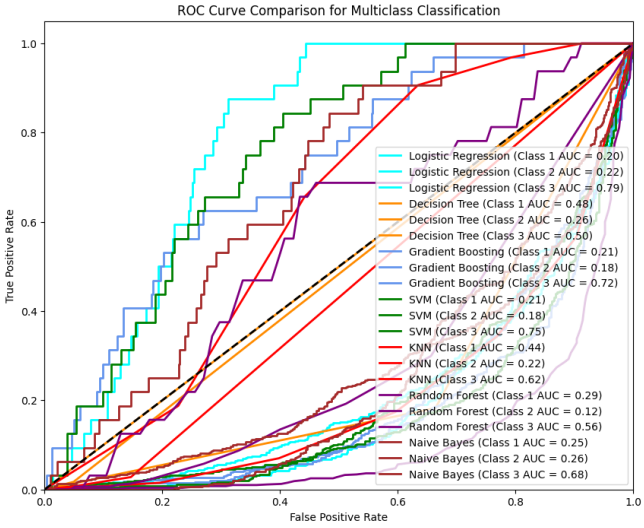


Fig:11 Distribution of Wine Quality for Random Forest

- The 'medium' quality category is the most common.
- There are fewer 'low' quality wines than 'medium' quality wines.
- 'High' quality wines are the least common in the dataset.

This imbalance in the dataset explains why our model performs differently for each category. It's particularly good at identifying medium quality wines (the most common category) but struggles with high quality wines (the least common category).

3. **ROC Curve Comparision for Multiclass Classification**



- Class 1 (low): corresponds to wines with quality labeled as 'low' (quality $\leq$ 5). It refers to the low quality wines.
- Class 2 (medium): corresponds to wines with quality labeled as 'medium' (quality between 6 and 7). It refers to the medium quality wines.
- Class 3 (high): corresponds to wines with quality labeled as 'high' (quality > 7). It refers to the high quality wines.

IV. DISCUSSION

A. Limitations of approach:

Imbalance of data: The dataset consists of a significantly large number of medium category wines which might lead to biases. If there are more low-quality and medium-quality wines, the model can achieve even higher accuracy.

Feature selection: The features included in the dataset were according to availability and many relevant features could be missing. If there were other characteristics included in the data, the predicting power of the model could be enhanced.

Model complexity: Complex models like Random Forest can provide high accuracy but also lead to overfitting when training data is not large enough while simpler model might not capture patterns in the data perfectly.

Evaluation: The various metrics like accuracy, precision, and recall give a better understanding of performance but cannot be explained properly in a multi-class environment. More comprehensive metrics like the correlation coefficient could give better insights into imbalanced data.

Less hyperparameter tuning: The models were trained on default parameters with low tuning. Complex methods like

Grid search or random search could explain better performance.

B. Future improvements

- Balancing the data
- Feature engineering and exploring additional features
- Comprehensive model evaluation with wide range of evaluation metrics
- Better hyperparameter tuning
- Increase the data diversity
- Analyze other machine learning algorithms like ensemble methods and deep learning models to enhancing accuracy.
- Explore other techniques like regression method with different approach

V. *CONCLUSION:*

In this paper, we were able to analyze various classification model to predict the performance of wine quality based on physicochemical properties. We were able to achieve an accuracy of about 83% for Random Forest model making it a high performing model followed by gradient boosting model with accuracy of 76%. By categorizing the wines into different categories(low, medium, high) proves beneficial by simplifying the prediction task and increasing evaluation performance. Scaling also contributed to model performance and confusion matrix provides comprehensive description of performance.

Overall, this project gave us in-depth understanding of predictive modelling and classification that led to foundation for future improvement using advanced machine learning techniques. These techniques can help wine makers or producers make better decisions during wine production and selection. Also help brand their products accordingly in the market to capture the different market segments with different grades of wines. Future work should focus on better quality and additional important features that could contribute to enhancing the quality and what makes the Wine a best quality for consumers.

VI. REFERENCES

- *Piyush Bhardwaj, Parul Tiwari, Kenneth Olejar Jr, Wendy Parr , Don Kulasir. Vol 8, 15 June 2022, 100261. A machine learning application in wine quality prediction. https://www.sciencedirect.com/science/article/pii/S266682702200007X*

- *Dahal K.R., Dahal J.N., Banjade H., Gaire S. Jan (2021). Prediction of wine quality using machine learning algorithms.*

  Open Journal of Statistics 11(02):278-289. 10.4236/ojs.2021.112015

- *Yogesh Gupta. Volume 125, 2018. Selection of important features and predicting wine quality using machine learning techniques. https://www.sciencedirect.com/science/article/pii/S187705091732805*

- *The Dataset is been taken from UCI and Kaggle: https://archive.ics.uci.edu/dataset/186/wine+quality*

  https://www.kaggle.com/code/yasserh/wine-quality-prediction-comparing-top-ml-models/notebook

- *Machine Learning Using Python. Author: Mighty Itauma Itauma, PhD, Published August 16, 2024 https://amightyo.quarto.pub/machine-learning-using-python/*

VII. Conference Alignment

IWPR--EI 2025 : 2025 10th International Workshop on Pattern Recognition (IWPR 2025)
http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=182462&copyownerid=13881

When	Jun 13, 2025 - Jun 15, 2025
Where	Singapore
Submission Deadline	Jan 10, 2025

Conference Guidelines

- IWPR 2025 invites researchers and engineers from all related disciplines to submit research papers and special sessions proposals. All submissions are peer reviewed.
- Papers must be written in English.
- Up to 5 pages (including figures and references) are allowed for each paper. Up to five (5) extra pages are permissible with an additional fee (US$50/per page) paid at registration time.
- Authors should only submit original work that has neither appeared elsewhere for publication, nor is under review for another refereed publication. The

conference will verify the originality of manuscripts by comparing them with millions of other articles in databases worldwide.